Kap 13: Den Multiple Lineære Regressionsmodel: Modelkontrol.

25. maj 2002

Nr.	Hypotese	(A) Grafisk	(B) Numerisk	(C) Uddybende
-	**Modellen**	**$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$** Opstil evt. m. opgaverelevante termer , eller omdøb variable, eks.: løn = x_1, Uddannelse = x_2 og profit = y		
-	**Std.forudsætninger**	**I)** $E(e_i) = 0$ **II)** $Var(e_i) = \sigma^2$ for i =1,...,n. **III)** $Cov(e_i, e_j) = 0$ for $i \neq j \Rightarrow$ Uafhængige restled $\Rightarrow e_i \sim Nf(0,\sigma^2)$ ...**IV)** hvid støj		
-	**Standardiserede residualer**	$e_i^* = \dfrac{e_i}{\sqrt{Var(e_i)}}$, hvor var (e_i) afhænger af værdien af den forklarende variabel, mens var(e_i^*) er konstant (homoskedasticitet) for alle i = 1,...,n, så de kan sammenlignes direkte!		
1	Ad. **I)** Linearitet	- Plot $(y,\hat{y})$ -dia.: skal ligge om 45° linie. - Plot (x_i,e_i^*) – kontroller systematik	$R^2 = \dfrac{SAK_y - SRK}{SAK_y} > 0{,}5 \Rightarrow$ relativt høj	R^2= Determinationskoefficent, angiver *den andel af den totale variation i responsvariablen (y), der kan beskrives udfra den lineære relation til de forklarende variabel (x_j)*
2	Ad. **IV)** Normalfordelte std. residualer	Plot fraktil-dia. (Prob Plot): • Skal ligge *tilfældigt* om en ret linie		
3	Ad. II) Homoskedasticitet	- Plot (x_j,e_i^*): skal ligge usystematisk i et bredt bånd, uden for mange punkter uden for $(x_j,\|2,0\|)$		
4	Ad. **III)** Uafhængige restled (Autokorr.-kontrol) *NB: Kun ved **tidsrække** obs.*	- Plot (t,e_i^*): Kommenter passede antal fortegnsskift“	$DW = \dfrac{\sum_{i=2}^{n}(\hat{e}_i - \hat{e}_{i-1})^2}{\sum_{i=1}^{n}\hat{e}_i^2}$ 0 d_1 d_2 d_3 d_4 4 P A | ID | IA | ID | NA PA/NA=Pos./Neg. autoko r. ID=Indecisionsområdet IA = Ingen autokorrelati Forklar hvorfor PA/NA(hvis intuitivt)	- (t, ,e_i^*)= bløde kurver $\Rightarrow$ pos. autokorr. - (t, ,e_i^*) =ZIG-ZAG $\Rightarrow$ neg. autokorr. - DW kontrollerer for AR(1)-korrelation: dvs., om $e_i = \rho e_{i-1} + v_i$, *hvor* v_i er uaf. NF restled $\sim NF(0,\sigma^2)$ $Konf_{1-\alpha}[E(y)\|x_{i0})] = y \pm t_{1-\alpha/2}(n-p-1)\cdot s\sqrt{w(x)},\ s\sqrt{w(x)} = v\hat{a}r(\hat{y}\,\|\,x_{i0})$ $Pred_{1-\alpha}[(y)\|x_{i0})] = y \pm t_{1-\alpha/2}(n-p-1)\cdot s\sqrt{w(x)},\ s\sqrt{1+w(x)}$
5	Betydning af enkelt-obs.	- Plot (x_j,y)-dia.: • Kommenter om nogle obs. ligger ”isoleret” og dermed påvirker estimationen meget • Kommenter hvorledes de påvirker parameter estimaterne.	Cook's afstand: $D_i = \dfrac{(\hat{\beta}-\hat{\beta}_{(i)})^T(\vec{X}^T\vec{X})(\hat{\beta}-\hat{\beta}_{(i)})}{(p+1)s^2}$, $\hat{\beta}_{(i)}$ er MK parameterestimatet uden obs. nr. i.	- D_i = forskellen i parameterestimat med og uden obs. nr. i. - Anvendes til at udpege obs., der bør underkastes nærmere undersøgelse 1. Udpeg størst Cook’s D. 2. Sig hvorfor: Stor residual, ekstrem x-kombination eller begge. 3. Kommenter M_i (=Mahalanobis): intra- vs. ekstrapolation/yderkant af obs.
6	Multikollinaritet:	- Plot (x_{1i},x_{2i}) og kontroller for systematik	- $\rho = \dfrac{cov(X_1,X_2)}{\sqrt{var(X_1)\cdot var(X_2)}}$ (viser kun lineær korr.)	Def: der eksisterer et eller flere lineære bånd af typen: $c_o + c_1x_{1i} + \ldots + c_px_{pi} \approx 0$ *Konsekvens*: Stor Var$(\beta_i) \Rightarrow$ fortolkningsproblemer af parametre /effekten af enkelte forklarende variable på responsvariablen og brede konf/pred. *Symptomer:* 1) Lave værdier af partielle t-str. ses med stor R^2. 2) Parameterskøn har ikke forventet fortegn eller er insignifikante imod øko.teori. 3) Udeladelse af enkelt obs. eller enkelte forklarende variable ændrer parameterskøn meget.
7	H_0: $\beta_1=\beta_2=\ldots=\beta_p$		$V = \dfrac{R^2/p}{(1-R^2)/(n-p-1)} \sim F_{0,95}(n,n-p-1)$	
8	H_0: $\beta_i = \beta_{i0}$ (kan også være tosiddet)		$T_i = \dfrac{\hat{\beta}_i - \beta_{i0}}{\sqrt{var(\hat{\beta}_i)}} \sim t(n-p-1)$	